

The open-faced sandwich adjustment for MCMC using estimating functions

Benjamin A. Shaby

March 2, 2013

Abstract

The situation frequently arises where working with the likelihood function is problematic. This can happen for several reasons—perhaps the likelihood is prohibitively computationally expensive, perhaps it lacks some robustness property, or perhaps it is simply not known for the model under consideration. In these cases, it is often possible to specify alternative functions of the parameters and the data that can be maximized to obtain asymptotically normal estimates. However, these scenarios present obvious problems if one is interested in applying Bayesian techniques. Here we describe open-faced sandwich adjustment, a way to incorporate a wide class of non-likelihood objective functions within Bayesian-like models to obtain asymptotically valid parameter estimates and inference via MCMC. Two simulation examples show that the method provides accurate frequentist uncertainty estimates. The open-faced sandwich adjustment is applied to a Poisson spatio-temporal model to analyze an ornithology dataset from the citizen science initiative eBird.

1 Introduction

For many models arising in various fields of statistical analysis, working with the likelihood function can be undesirable. This may be the case for several reasons—perhaps the likelihood is prohibitively expensive to compute, perhaps it presumes knowledge of a component of the model that one is unwilling to specify, or perhaps its form is not even known for a chosen probability model. Such scenarios present problems if one wishes to perform Bayesian analysis. Applying the Bayesian computational and inferential machinery, thereby enjoying benefits such as natural shrinkage,

1 variance propagation, and the ability to incorporate complex hierarchical
 2 dependences, usually requires working directly with the likelihood function.
 3 To motivate the development, we briefly describe the analysis of bird
 4 sightings contained in Section 5.1. The data consist of several thousand
 5 counts occurring irregularly in space and time (see Figure 1), along with
 6 several spatially-varying covariates carefully chosen by a group of ornithol-
 7 ogists. A natural model for such data is a hierarchical Poisson regression
 8 with a random effect specified as a spatio-temporal Gaussian process with
 9 unknown covariance parameters. Here, whatever one’s philosophical orien-
 10 tation, Bayesian methods are most practical to implement, and in addition
 11 provide sharing of information across space and time, as well as automatic
 12 uncertainty estimation of predictive abundance maps. Furthermore, ob-
 13 taining an MCMC sample of the posterior distribution is desirable because
 14 inferences on the posterior correlation surface of the random effect, a nonlin-
 15 ear functional of random covariance parameters, is of independent interest
 16 to the ornithologists. However, the sheer size of the dataset makes MCMC
 17 under this model intractable, so a faster objective function is used in place of
 18 the high-dimensional Gaussian likelihood. The goal of the method presented
 19 here is to enable such a substitution while retaining a valid interpretation
 20 of the resultant MCMC sample.

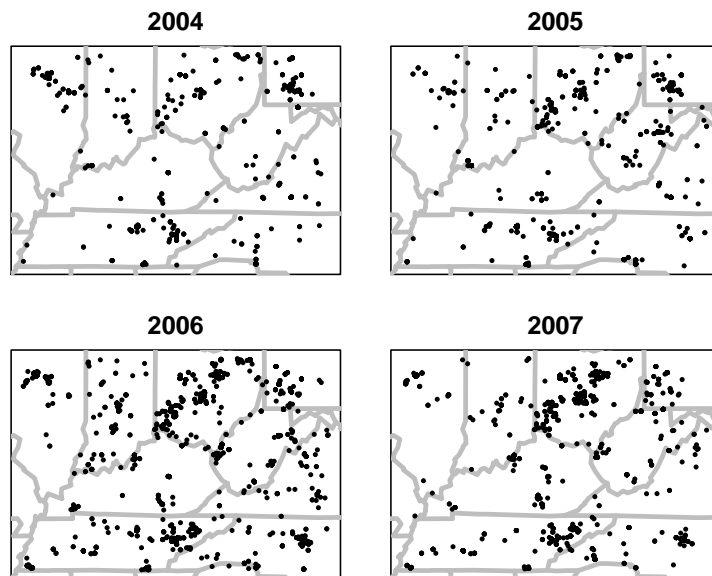


Figure 1: Spatial locations of Northern Cardinal observations.

1 More generally, suppose that one specifies a model, either one-stage or
2 hierarchical, and wants the advantages of being Bayesian, but the likelihood
3 in some level of the hierarchy is problematic. Suppose, however, that one
4 can write down some objective function $\ell_M(\boldsymbol{\theta}; \mathbf{y})$ of the parameters and the
5 data (possibly conditional on other parameters) that behaves similarly to
6 the log likelihood. We will define what we mean by “similarly” in Section
7 2. Important examples of methods that employ such objective functions
8 include generalized estimating equations (Hardin and Hilbe, 2003), general-
9 ized method of moments (Hall, 2005), and robust M-estimation (Huber and
10 Ronchetti, 2009), as well as the two examples we will consider here, covari-
11 ance tapering (Kaufman et al., 2008) and composite likelihoods (Lindsay,
12 1988).

13 The question we attempt to answer here is this: Can we insert $\ell_M(\boldsymbol{\theta}; \mathbf{y})$,
14 in place of the likelihood, into an MCMC algorithm like Metropolis-Hastings
15 and “trick” it into doing something useful? We claim that we can—that for
16 many useful examples, simply swapping $\ell_M(\boldsymbol{\theta}; \mathbf{y})$ into a sampler results in
17 a *quasi*-posterior sample that can be rotated and scaled to yield desirable
18 properties.

19 The OFS adjustment relies on asymptotic theory that was formally devel-
20 oped in Chernozhukov and Hong (2003), but which is quite intuitive. These
21 authors were interested in using Metropolis-Hastings as an optimization al-
22 gorithm for badly-behaved objective functions, not in using non-likelihood
23 objective functions for performing Bayesian-like analysis, as we are here.
24 Although their goals were entirely different, the theory contained therein is
25 extremely useful for our purposes.

26 Previous attempts to incorporate non-likelihood objective functions into
27 the Bayesian setting, to our knowledge, have been few. McVean et al. (2004)
28 use composite likelihoods within reversible jump MCMC, without any ad-
29 justment, to estimate population genetic parameters. Realizing that their
30 sampler would result in invalid inferences, McVean et al. (2004) turn to a
31 parametric bootstrap to estimate sampling variability. Smith and Stephen-
32 son (2009) were interested in max-stable processes for spatial extreme value
33 analysis. They also use composite likelihoods within MCMC without ad-
34 justment. The special case of using the generalized method of moments
35 objective function (Hansen, 1982; Hall, 2005) for generalized linear models
36 within an MCMC sampler was explored by Yin (2009). Tangentially related
37 is Tian et al. (2007), who use MCMC to estimate the sampling distribution
38 of $\hat{\boldsymbol{\theta}} = \operatorname{argmax} \ell_M(\boldsymbol{\theta}; \mathbf{y})$.

39 Cooley et al. (2012) attempt to solve the same problem that we address
40 here. Whereas we adjust quasi-posterior samples generated from MCMC

1 post hoc, these authors propose an adjustment to the Metropolis likelihood
2 ratio within the sampler itself. Their goal, like ours, is to achieve desir-
3 able frequentist coverage properties of credible intervals computed based on
4 MCMC. Although their approach is quite general, Cooley et al. (2012) re-
5 strict their attention to using composite likelihoods for max-stable processes.
6 The approach taken in the present article is closely related to that of Cooley
7 et al. (2012), but the OFS adjustment differs from their adjustment in its
8 structure as well as its motivating asymptotic arguments.

9 Both the motivating insights for the OFS adjustment and the criterion
10 by which we evaluate it is essentially the idea of calibration (Draper, 2006).
11 In our interpretation, a well-calibrated method has the property that when
12 used to construct credible intervals from many different datasets, those in-
13 tervals ought to cover the true parameter at close to their nominal rates.
14 Essentially, this says that well-calibrated credible intervals behave like con-
15 fidence intervals. If we construct intervals with accurate coverage directly as
16 the $\alpha/2$ and $(1 - \alpha/2)$ empirical quantiles of an MCMC sample for different
17 values of α , we claim that in some way our uncertainty about a parameter is
18 well-described by the sample. Evaluating an approximate Bayesian method
19 by this criterion has intuitive practical appeal, and it has been endorsed in
20 particular by objective Bayesians (Bayarri and Berger, 2004; Berger et al.,
21 2001, e.g.).

22 This principle, along with some basic asymptotic observations, leads to
23 the OFS adjustment. The asymptotic theory gives us the limiting normal
24 distribution of quasi-Bayes point estimators. We take this distribution, in
25 an informal sense, to be a summary of our uncertainty about θ , up to
26 an asymptotic approximation. The asymptotic theory also gives us the
27 limiting normal distribution of the quasi-posterior. Since these two limiting
28 distributions are not, in general, the same, and since we would like the
29 quasi-posterior to summarize our uncertainty about θ in the sense of being
30 well-calibrated, our strategy is to adjust samples from the quasi-posterior
31 so that their limiting distribution matches that of the quasi-Bayesian point
32 estimator.

33 We note the temptation to ask how well the adjusted quasi-posterior
34 distribution approximates the true posterior distribution, in cases when the
35 true likelihood is available. However, this is the incorrect comparison to
36 make. The true posterior distribution contains the information about θ
37 obtained through the likelihood. When some other function ℓ_M is used in
38 place of the likelihood, there is no reason to expect the information content
39 to remain the same. We would like the adjusted quasi-posterior distribution
40 to represent this loss of information, not hide it. In our simulation examples

1 in Section 4, the frequentist accuracy of credible intervals based on adjusted
2 quasi-posterior samples shows that the OFS adjustment accomplishes this
3 task.

4 Throughout, it will be assumed that expectations will be computed with
5 respect to the true parameter $\boldsymbol{\theta}_0$. We define the square root of a symmetric
6 positive definite matrix \mathbf{A} to be $\mathbf{A}^{1/2} = \mathbf{O}\mathbf{D}^{1/2}\mathbf{O}'$, where $\mathbf{A} = \mathbf{O}\mathbf{D}\mathbf{O}'$ with
7 \mathbf{O} orthogonal and \mathbf{D} diagonal. The square root of a matrix is not unique;
8 here we compute $\mathbf{A}^{1/2}$ using the singular value decomposition, which is
9 numerically stable and preserves key geometric attributes.

10 We begin in Section 2 by defining the quasi-Bayesian framework and
11 reviewing the relevant asymptotic theory. In Section 3 we develop the OFS
12 adjustment method, and we demonstrate how to apply it in two different
13 statistical contexts in Section 4. In Section 5 we apply the OFS adjustment
14 to analyze a dataset of Northern Cardinal sightings taken from the citizen
15 science project eBird. Section 6 concludes.

16 2 The quasi-Bayesian framework

17 We begin by assuming that the parameter of interest $\boldsymbol{\theta}$ lies in the interior
18 of a compact convex space $\boldsymbol{\Theta}$. Suppose we are given \mathbf{y} , which consists of
19 n observations, from which we wish to estimate $\boldsymbol{\theta}$. Suppose further that
20 we have at our disposal some objective function $\ell_M(\boldsymbol{\theta}; \mathbf{y})$ from which it is
21 possible to compute $\hat{\boldsymbol{\theta}}_M = \operatorname{argmax}_{\boldsymbol{\Theta}} \ell_M(\boldsymbol{\theta}; \mathbf{y})$.

22 Following Chernozhukov and Hong (2003), we define the *quasi-posterior*
23 distribution based on n observations as

$$\pi_{M,n}(\boldsymbol{\theta}|\mathbf{y}_n) = \frac{L_{M,n}(\boldsymbol{\theta}; \mathbf{y}_n)\pi(\boldsymbol{\theta})}{\int_{\boldsymbol{\Theta}} L_{M,n}(\boldsymbol{\theta}; \mathbf{y}_n)\pi(\boldsymbol{\theta}) d\boldsymbol{\theta}}, \quad (1)$$

24 where $L_{M,n}(\boldsymbol{\theta}; \mathbf{y}_n) = \exp\{\ell_{M,n}(\boldsymbol{\theta}; \mathbf{y}_n)\}$, and $\pi(\boldsymbol{\theta})$ is a prior density on $\boldsymbol{\theta}$. We
25 will assume, for convenience, that $\pi(\boldsymbol{\theta})$ is proper with support on $\boldsymbol{\Theta}$. The
26 function $L_{M,n}$ is *not* necessarily a density, and thus $\pi_{M,n}(\boldsymbol{\theta}|\mathbf{y}_n)$ is not a true
27 posterior density in any probabilistic sense. We will assume, however, that
28 $L_{M,n}$ is integrable, so as long as the prior $\pi(\boldsymbol{\theta})$ is proper, it easily follows
29 that $\pi_{M,n}(\boldsymbol{\theta}|\mathbf{y}_n)$ will be a proper density.

30 Equipped with notion of a quasi-posterior density, we can define quasi-
31 posterior risk as $R_n(\boldsymbol{\theta}) = \int_{\boldsymbol{\Theta}} \rho_n(\boldsymbol{\theta} - \boldsymbol{\theta}^*)\pi_{M,n}(\boldsymbol{\theta}^*|\mathbf{y}_n) d\boldsymbol{\theta}^*$, where $\rho_n(\mathbf{u})$ is
32 some convex scalar loss function. For simplicity, we assume that $\rho_n(\mathbf{u})$ is
33 symmetric, although this assumption may be dropped. Then for a given
34 loss function, the quasi-Bayes estimator is naturally defined as $\hat{\boldsymbol{\theta}}_{\text{QB}} =$
35 $\operatorname{argmin}_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} R_n(\boldsymbol{\theta})$, the value of $\boldsymbol{\theta}$ that minimizes quasi-posterior risk.

Our requirements on $\ell_{M,n}(\boldsymbol{\theta}; \mathbf{y}_n)$ are fairly minimal and are met by most objective functions in wide use in statistics. Technical assumptions are contained in Chernozhukov and Hong (2003), but they are in general satisfied when $\hat{\boldsymbol{\theta}}_M$ is weakly consistent for $\boldsymbol{\theta}_0$ and asymptotically normal. Asymptotic normality of $\hat{\boldsymbol{\theta}}_M$ is of the form

$$\mathbf{J}_n^{1/2}(\hat{\boldsymbol{\theta}}_{M,n} - \boldsymbol{\theta}_0) \xrightarrow{\mathcal{D}} N(\mathbf{0}, \mathbf{I}), \quad (2)$$

where

$$\begin{aligned} \mathbf{J}_n &= \mathbf{Q}_n \mathbf{P}_n^{-1} \mathbf{Q}_n \\ \mathbf{P}_n &= E_0[\nabla_0 \ell_{M,n} \nabla_0 \ell'_{M,n}] \\ \mathbf{Q}_n &= -E_0[\mathcal{H}_0 \ell_{M,n}]. \end{aligned} \quad (3)$$

The notation $\nabla_0 f$ refers to the gradient of the function f evaluated at the true parameter $\boldsymbol{\theta}_0$, and $\mathcal{H}_0 f$ refers to the Hessian of f evaluated at $\boldsymbol{\theta}_0$. These matrices have been defined in terms of partial derivatives, but in general, $\ell_{M,n}$ does not have to be differentiable or even continuous for the theory to apply. In this case, small adjustments of the definitions of \mathbf{P}_n and \mathbf{Q}_n are necessary.

The sandwich matrix \mathbf{J}_n^{-1} is familiar from generalized estimating equations, quasi-likelihood, and other areas, and is referred to by various names, including the *Godambe information criterion* and the *robust information criterion* (e.g. Durbin, 1960; Bhapkar, 1972; Morton, 1981; Ferreira, 1982; Godambe and Heyde, 1987; Heyde, 1997). We note that in the special case when $\ell_{M,n}(\boldsymbol{\theta}; \mathbf{y})$ is the true likelihood, $\mathbf{Q}_n \equiv \mathbf{J}_n$, the Fisher information. We will hereafter assume that this is not the case.

2.1 Review of relevant asymptotic theory

Chernozhukov and Hong (2003) elucidates the asymptotic behavior of $\pi_{M,n}(\boldsymbol{\theta}|\mathbf{y}_n)$, which motivates the open-face sandwich adjustment. These results are direct analogues of well-known asymptotic properties of true posterior distributions. Their Theorem 2, which we re-state below, states that the asymptotic distribution of the quasi-Bayes estimator $\hat{\boldsymbol{\theta}}_{QB,n}$ is the same as that of the extremum estimator $\hat{\boldsymbol{\theta}}_{M,n}$.

Theorem 1 *Assuming sufficient regularity of $\ell_{M,n}(\boldsymbol{\theta}; \mathbf{y}_n)$,*

$$\mathbf{J}_n^{1/2}(\hat{\boldsymbol{\theta}}_{QB} - \boldsymbol{\theta}_0) \xrightarrow{\mathcal{D}} N(\mathbf{0}, \mathbf{I}).$$

1 Theorem 1 above is the quasi-posterior extension of the well-known result
2 that, under fairly general conditions, Bayesian point estimates have the same
3 asymptotic distribution as maximum likelihood estimates.

4 Theorem 1 of Chernozhukov and Hong (2003), which we re-state here
5 in a slightly different form, is a kind of quasi-Bayesian consistency result,
6 showing that quasi-posterior mass accumulates at the true parameter θ_0 .

7 **Theorem 2** *Under the same conditions as Theorem 1*

$$\|\pi_{M,n}(\theta|\mathbf{y}_n) - \pi_{M,\infty}(\theta|\mathbf{y}_n)\|_{TV} \xrightarrow{\mathcal{P}} 0,$$

8 where $\|\cdot\|_{TV}$ indicates the total variation norm, and $\pi_{M,\infty}(\theta|\mathbf{y}_n)$ is a normal
9 density with random mean $\theta_0 + \mathbf{Q}_n^{-1}\nabla\ell_{M,n}(\theta_0)$ and covariance matrix \mathbf{Q}_n^{-1} .

10 Theorem 2 may be arrived at informally via a simple Taylor series argument.
11 It is therefore intuitive that the quasi-posterior converges to limiting normal
12 distribution whose covariance matrix is defined by the second derivatives of
13 ℓ_M .

14 The key observation is that the limiting quasi-posterior distribution has
15 a *different* covariance matrix than the asymptotic sampling distribution of
16 the quasi-Bayes point estimate. The consequence is that the usual Bayesian
17 method of constructing credible intervals based on quantiles of the quasi-
18 posterior sample will, viewed as confidence intervals, not have their nominal
19 frequentist coverage probabilities. Fortunately, thanks to Chernozhukov and
20 Hong (2003), we know what those two asymptotic covariance matrices look
21 like, which suggests a way to “fix” $\pi_{M,n}(\theta|\mathbf{y}_n)$.

22 3 The open-faced sandwich adjustment

23 Let us assume that we have a sample of draws from $\pi_{M,n}(\theta|\mathbf{y}_n)$, generated
24 by replacing the likelihood with $\ell_M(\theta; \mathbf{y})$ in some MCMC sampler such as
25 Metropolis-Hastings. Our aim here is to adjust the quasi-posterior draws
26 such that the adjusted sample realistically reflects how the data informs our
27 uncertainty about the parameter of interest θ through the function $\ell_M(\theta; \mathbf{y})$.
28 Were that the case, the usual credible intervals constructed from empirical
29 quantiles of the adjusted sample would have close to nominal coverage. We
30 will accomplish this by constructing a matrix Ω_n that, when applied to
31 the (centered) quasi-posterior sample, will rotate and scale the points in an
32 appropriate way.

1 We have observed that whereas the asymptotic covariance matrix of $\hat{\boldsymbol{\theta}}_{M,n}$
2 is the sandwich matrix \mathbf{J}_n^{-1} , the asymptotic covariance matrix of the quasi-
3 posterior distribution is a single “slice of bread” \mathbf{Q}_n^{-1} . What we want to
4 do then is complete the sandwich by joining the slice of bread \mathbf{Q}_n^{-1} to the
5 open-faced sandwich $\mathbf{P}_n \mathbf{Q}_n^{-1}$ to get \mathbf{J}_n^{-1} .

6 We define $\boldsymbol{\Omega}_n = \mathbf{Q}_n^{-1} \mathbf{P}_n^{1/2} \mathbf{Q}_n^{1/2}$, the open-faced sandwich adjustment ma-
7 trix. One can easily check that if $\mathbf{Z}_n \sim N(\mathbf{0}, \mathbf{Q}_n^{-1})$, then $\boldsymbol{\Omega}_n \mathbf{Z}_n \sim N(\mathbf{0}, \mathbf{J}_n^{-1})$.
8 The idea then is to take samples from $\pi_M(\boldsymbol{\theta}|\mathbf{y})$ obtained via MCMC and
9 pre-multiply them (after centering) by an estimator $\hat{\boldsymbol{\Omega}}$ of $\boldsymbol{\Omega}$ to “correct” the
10 quasi-posterior sample. That is, if $\boldsymbol{\theta}^{(1)}, \dots, \boldsymbol{\theta}^{(J)}$ is a sample from $\pi_M(\boldsymbol{\theta}|\mathbf{y})$,
11 then for each $j = 1, \dots, J$,

$$\boldsymbol{\theta}_{\text{OFS}}^{(j)} = \hat{\boldsymbol{\theta}}_{\text{QB}} + \hat{\boldsymbol{\Omega}}(\boldsymbol{\theta}^{(j)} - \hat{\boldsymbol{\theta}}_{\text{QB}}) \quad (4)$$

12 is the open-face sandwich adjusted sample. It is clear that a consistent
13 estimator of $\boldsymbol{\Omega}$ will generate credible intervals that are consistent $(1 - \alpha)$
14 confidence intervals.

15 3.1 Estimating $\boldsymbol{\Omega}$

16 The OFS adjustment (4) requires an estimate of the matrix $\boldsymbol{\Omega}$, which in
17 turn requires estimates of \mathbf{P} and \mathbf{Q} . Because the OFS adjustment occurs
18 post-hoc, it is possible to leverage the existing MCMC sample to compute
19 $\hat{\boldsymbol{\Omega}}$. There are many possible approaches to this task, and here we offer some
20 suggestions, which we summarize in Table 1.

21 While \mathbf{P} is notoriously difficult to estimate well (see Kauermann and
22 Carroll, 2001, for some examples), Theorem 1 immediately suggests a way
23 to estimate \mathbf{Q} directly from the MCMC sample with almost no additional
24 computational cost. Specifically, noting that the quasi-posterior density con-
25 verges to a normal with covariance matrix \mathbf{Q}^{-1} , a natural estimate $\hat{\mathbf{Q}}_{\text{I}}^{-1}$ is
26 just the sample covariance matrix of the MCMC sample. Another possibil-
27 ity that requires almost no additional computation is to retain the results of
28 the evaluations of ℓ_M at each iteration of the sampler and use them to nu-
29 merically estimate the Hessian matrix at $\hat{\boldsymbol{\theta}}_{\text{QB}}$. This Hessian approximation
30 will generally be a good estimator $\hat{\mathbf{Q}}_{\text{II}}$ of \mathbf{Q} .

31 These estimators of \mathbf{Q} are not only simple to compute, but they arise as
32 direct results of MCMC output, requiring no additional analytical deriva-
33 tions based on ℓ_M . They are, in this sense, “model-blind.” Unfortunately,
34 we are unaware of any such “model-blind” estimators of \mathbf{P} . The simplest
35 solution, in the case where we can write an expression for $\nabla \ell_M(\boldsymbol{\theta}; \mathbf{y})$ and the

1 data \mathbf{y} consists of n independent replicates, is to compute a basic moment
 2 estimator

$$\hat{\mathbf{P}}_{\text{I}} = \frac{1}{n} \sum_{i=1}^n \nabla \ell_M(\hat{\boldsymbol{\theta}}_{\text{QB}}; \mathbf{y}_i) \nabla \ell_M(\hat{\boldsymbol{\theta}}_{\text{QB}}; \mathbf{y}_i)', \quad (5)$$

3 which is consistent as $n \rightarrow \infty$ under standard regularity conditions. We use
 4 equation (5) in Section and 4.2, where we have replication. However, because
 5 $\nabla \ell_M(\hat{\boldsymbol{\theta}}_{\text{QB}}; \mathbf{y})$ converges to zero, when we only observe a single realization
 6 of a stochastic process, as in Section 4.1, equation (5) fails to provide a
 7 viable estimator. In this latter example, analytical expressions for $\mathbf{P}(\boldsymbol{\theta})$ are
 8 available. Plugging $\hat{\boldsymbol{\theta}}_{\text{QB}}$ into the analytical asymptotic expression gives an
 9 estimator $\hat{\mathbf{P}}_{\text{II}}$. If a corresponding analytical expression exists for \mathbf{Q} , we call
 10 the corresponding plug-in estimator $\hat{\mathbf{Q}}_{\text{III}}$

11 When an expression for $\mathbf{P}(\boldsymbol{\theta})$ is unavailable, but when it is possible to
 12 simulate the process that generated \mathbf{y} , the parametric bootstrap is an attrac-
 13 tive option. Let $\mathbf{y}_1, \dots, \mathbf{y}_K$ be K independent realizations of the stochastic
 14 process generated under $\hat{\boldsymbol{\theta}}_{\text{QB}}$. Then

$$\hat{\mathbf{P}}_{\text{boot}} = \frac{1}{K} \sum_{k=1}^K \nabla \ell_M(\hat{\boldsymbol{\theta}}_{\text{QB}}; \mathbf{y}_k) \nabla \ell_M(\hat{\boldsymbol{\theta}}_{\text{QB}}; \mathbf{y}_k)' \quad (6)$$

15 is the parametric bootstrap estimator of \mathbf{P} (an analogous estimator could,
 16 of course, be used for \mathbf{Q}). A nice feature of (5) and (6) is that, at the ex-
 17 pense of (perhaps considerable) computational effort, one could substitute
 18 finite-difference approximations to the required gradients to obtain reason-
 19 able estimators, even in the absence of available closed-form expressions for
 20 $\nabla \ell_M(\boldsymbol{\theta}; \mathbf{y})$.

Estimator	Description
$\hat{\mathbf{Q}}_{\text{I}}^{-1}$	sample covariance of MCMC sample
$\hat{\mathbf{Q}}_{\text{II}}$	Hessian of $\ell_M(\hat{\boldsymbol{\theta}}_{\text{QB}})$
$\hat{\mathbf{Q}}_{\text{III}}$	plug $\hat{\boldsymbol{\theta}}_{\text{QB}}$ into asymptotic formula
$\hat{\mathbf{P}}_{\text{I}}$	moment estimator based on score vector
$\hat{\mathbf{P}}_{\text{II}}$	plug $\hat{\boldsymbol{\theta}}_{\text{QB}}$ into asymptotic formula
$\hat{\mathbf{P}}_{\text{boot}}$	parametric bootstrap

Table 1: Summary of estimators of sandwich components.

1 3.2 The curvature adjustment

2 We now describe the curvature-adjusted sampler of Cooley et al. (2012).
 3 This sampler was presented as a way to include composite likelihoods in
 4 Bayesian-like models but in fact has far wider generality. Composite likeli-
 5 hoods (Lindsay, 1988) are functions of $\boldsymbol{\theta}$ and \mathbf{y} constructed as the product
 6 of joint densities of subsets of the data. In effect, composite likelihoods
 7 treat these subsets as though they were independent. Under fairly general
 8 regularity conditions, the asymptotic distribution of maximum composite
 9 likelihood estimators have sandwich form (2) (Lindsay, 1988). Although
 10 Cooley et al. (2012) consider only composite likelihoods, their argument
 11 holds equally well for any function $\ell_M(\boldsymbol{\theta}; y)$ with sandwich asymptotics.

12 The curvature-adjusted sampler begins by computing the extremum es-
 13 timator $\hat{\boldsymbol{\theta}}_M$ and $\hat{\boldsymbol{\Omega}}(\hat{\boldsymbol{\theta}}_M)$ as a preliminary step. It works by modifying the
 14 Metropolis-Hastings algorithm using a transformation of the form (4) such
 15 that the acceptance ratio has a desirable asymptotic distribution. Specifi-
 16 cally, at each iteration $j = 1, \dots, J$, the algorithm proposes a new value $\boldsymbol{\theta}^*$
 17 from some density $q(\cdot|\boldsymbol{\theta}^{(j)})$ and compares it to the current state $\boldsymbol{\theta}^{(j)}$ to eval-
 18 uate whether to accept or reject the proposal. The difference between the
 19 curvature-adjusted sampler and the traditional Metropolis-Hastings sam-
 20 pler is that this comparison now takes place after $\boldsymbol{\theta}^*$ and $\boldsymbol{\theta}^{(j)}$ are scaled and
 21 rotated. That is, $\boldsymbol{\theta}^*$ is accepted with probability

$$\min \left\{ 1, \frac{L_M(\boldsymbol{\theta}_{CA}^*; \mathbf{y}) \pi(\boldsymbol{\theta}^*) q(\boldsymbol{\theta}^{(j)} | \boldsymbol{\theta}^*)}{L_M(\boldsymbol{\theta}_{CA}^{(j)}; \mathbf{y}) \pi(\boldsymbol{\theta}^{(j)}) q(\boldsymbol{\theta}^* | \boldsymbol{\theta}^{(j)})} \right\} \quad (7)$$

22 where $\boldsymbol{\theta}_{CA}^* = \hat{\boldsymbol{\theta}}_M + \hat{\boldsymbol{\Omega}}(\hat{\boldsymbol{\theta}}_M)(\boldsymbol{\theta}^* - \hat{\boldsymbol{\theta}}_M)$, and analogously for $\boldsymbol{\theta}_{CA}^{(j)}$. Cooley et al.
 23 (2012) use a result from Kent (1982) to show that the ratio in (7) has the
 24 same asymptotic distribution as that of the true likelihood ratio, and argue
 25 that the resultant sample has an asymptotic stationary distribution that is
 26 normal with covariance \mathbf{J}^{-1} , as desired. Note that unlike the OFS adjust-
 27 ment, the curvature-adjusted sampler requires outside initial estimates of $\boldsymbol{\theta}$
 28 and $\boldsymbol{\Omega}$ because the adjustment occurs within the sampling algorithm.

29 3.3 OFS within a Gibbs sampler

30 With some care, the OFS adjustment may be applied in the Gibbs sampler
 31 setting. Suppose we divide $\boldsymbol{\theta}$ into B blocks such that $\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_B$ forms a par-
 32 tition of $\boldsymbol{\theta}$, and we wish to draw from the quasi-full conditional distribution
 33 with density $f(\boldsymbol{\theta}_i | \boldsymbol{\theta}_{-i}, \mathbf{y}) \propto L_M(\mathbf{y} | \boldsymbol{\theta}) f(\boldsymbol{\theta}_i)$, where $\boldsymbol{\theta}_{-i}$ refers to the elements
 34 of $\boldsymbol{\theta}$ not contained in $\boldsymbol{\theta}_i$. Then the adjustment matrix $\boldsymbol{\Omega}_{\boldsymbol{\theta}_i | \boldsymbol{\theta}_{-i}}$ is defined as

1 before, only now it applies only to θ_i and is conditional on θ_{-i} . If all full
2 conditional densities $f(\theta_i|\theta_{-i}, \mathbf{y})$, $i = 1, \dots, B$, are quasi-full conditional
3 densities in the sense that they are proportional to a product of $\ell_M(\mathbf{y}|\theta)$
4 and another density, the Gibbs sampler may be run by successively drawing
5 from $f(\theta_i|\theta_{-i}, \mathbf{y})$, $i = 1, \dots, B$, and the OFS adjustment may proceed post
6 hoc as before. This can be seen by viewing the Gibbs sampler as a special
7 case of Metropolis-Hastings (see Robert and Casella, 2004, Section 7.1.4).

8 Now suppose that at iteration j of a Gibbs sampler we have drawn $\theta_i^{(j)}$
9 from the quasi-full conditional $f(\theta_i|\theta_{-i}^{(j)}, \mathbf{y})$. Suppose further that $f(\theta_{i+1}|\theta_{-(i+1)}^{(j)}, \mathbf{y})$
10 is *not* a function of L_M , as will be the case for many parameters in hier-
11 archical models that contain L_M . Since $f(\theta_{i+1}|\theta_{-(i+1)}^{(j)}, \mathbf{y})$ is a true full
12 conditional density and not a quasi-full conditional density, there is no OFS
13 adjustment to make. But clearly $f(\theta_{i+1}|\theta_{-(i+1)}^{(j)}, \mathbf{y})$ depends on $\theta_i^{(j)}$, and
14 as a result, plugging in un-adjusted samples of $\theta_i^{(j)}$ will not result in the
15 desired stationary distribution. It is clear then that to achieve proper vari-
16 ance propagation through the model, we must adjust $\theta_i^{(j)}$ *before* plugging
17 it into $f(\theta_{i+1}|\theta_{-(i+1)}^{(j)}, \mathbf{y})$. Therefore, the OFS adjustment within the Gibbs
18 sampler may not be applied post hoc, but rather must occur within the
19 sampling algorithm.

20 Embedding OFS adjustments within Gibbs samplers requires careful
21 consideration of the conditional OFS matrices $\Omega_{\theta_i|\theta_{-i}}$, $i = 1, \dots, B$, the
22 adjustment matrices associated with the quasi-full conditional distributions
23 $f(\theta_i|\theta_{-i})$, $i = 1, \dots, B$. Because it is defined conditionally on θ_{-i} , ide-
24 ally each $\Omega_{\theta_i|\theta_{-i}}$ should be re-estimated at each iteration j based on the
25 current value $\theta_{-i}^{(j)}$. We refer to $\hat{\Omega}_{\theta_i|\theta_{-i}}^{(j)}$ as the conditional OFS adjustment
26 matrix for $\theta_i^{(j)}$ at iteration j . Implementing the OFS Gibbs sampler us-
27 ing these conditional OFS adjustments requires estimation of $\hat{\Omega}_{\theta_i|\theta_{-i}}^{(j)}$, by
28 one of the techniques described in Section 3.1, for each block of parameters
29 $i = 1, \dots, B$ with corresponding quasi-full conditional depending on L_M ,
30 and for each Gibbs iteration j in $1, \dots, J$. Furthermore, each computation
31 of $\hat{\Omega}_{\theta_i|\theta_{-i}}^{(j)}$ requires an estimate of $\theta_i^{(j)}$ as input, necessitating some sort of
32 optimization or MCMC, again nested within each block $i = 1, \dots, B$ and
33 each iteration j in $1, \dots, J$. This is an enormous computational burden.

34 Instead, we make the simplifying assumption that $\Omega_{\theta_i|\theta_{-i}}^{(j)}$ does not change
35 much from iteration to iteration. We instead use a constant (with respect to
36 j) estimate $\hat{\Omega}_{\theta_i|\hat{\theta}_{-i}}$, where $\hat{\theta}_{-i}$ is fixed at its marginal quasi-Bayes estimate.
37 We refer to $\hat{\Omega}_{\theta_i|\hat{\theta}_{-i}}$ as the marginal OFS adjustment matrix for $\theta_i^{(j)}$.

1 Despite this simplification, the OFS-adjusted Gibbs sampler still requires
 2 additional work relative to an un-adjusted sampler because the algorithm
 3 requires $\hat{\boldsymbol{\theta}}_{\text{QB}}$ and $\hat{\boldsymbol{\Omega}}_{\boldsymbol{\theta}_i|\hat{\boldsymbol{\theta}}_{-i}}$, $i = 1, \dots, B$, as input. In practice, then, we
 4 run the sampler twice. The first time, recalling that Theorem 2 says the
 5 un-adjusted quasi-posterior concentrates its mass at $\boldsymbol{\theta}_0$, we make no OFS
 6 adjustments, and use the generated sample to produce $\hat{\boldsymbol{\theta}}_{\text{QB}}$. We next use
 7 $\hat{\boldsymbol{\theta}}_{\text{QB}}$ to produce $\hat{\boldsymbol{\Omega}}_{\boldsymbol{\theta}_i|\hat{\boldsymbol{\theta}}_{-i}}$, $i = 1, \dots, B$, using one of the methods described
 8 in Section 3.1. Finally, the Gibbs sampler is re-run, this time with the
 9 transformation defined by (4) applied for each block i and each iteration j ,
 10 $i = 1, \dots, B$, $j = 1, \dots, J$. Thus, the computational burden required to use
 11 marginal OFS adjustments is approximately twice that of the un-adjusted
 12 Gibbs sampler. In contrast, the additional computational burden required
 13 to use conditional OFS adjustments may range from several fold to several
 14 thousand fold, depending on the method used to estimate $\boldsymbol{\Omega}_{\boldsymbol{\theta}_i|\boldsymbol{\theta}_{-i}}^{(j)}$.

15 We have explored (informally) the effects of using the much more com-
 16 putationally efficient marginal adjustment instead of the conditional ad-
 17 justment and found only very minor differences in the resultant adjusted
 18 quasi-posteriors. (See Section 4.1 for an example.) The issue of conditional
 19 vs. marginal adjustments also appears in Cooley et al. (2012). They refer
 20 to using constant (in j) adjustment matrices as an “overall” Gibbs sampler
 21 and using conditional adjustment matrices as an “adaptive” Gibbs sampler.
 22 Corroborating our findings, Cooley et al. (2012), in a more systematic study
 23 using a very simple model, found very little difference between their overall
 24 and adaptive curvature-adjusted quasi-posteriors.

25 4 Examples

26 We now describe two examples of non-likelihood objective functions that
 27 have appeared in the literature. In each example, working with the likeli-
 28 hood is problematic for a different reason, and each fits into the OFS frame-
 29 work. In the first example, we apply covariance tapering (Furrer et al., 2006;
 30 Kaufman et al., 2008; Shaby and Ruppert, 2012) to large spatial datasets.
 31 Here the likelihood requires the numerical inversion of a very large matrix.
 32 For large datasets, this inversion becomes prohibitively computationally-
 33 expensive, so the likelihood is replaced with its tapered version, which lever-
 34 ages sparse-matrix algorithms to speed up computations. In the second ex-
 35 ample, composite likelihoods for spatial max-stable processes (Smith, 1990;
 36 Padoan et al., 2010), a probability model is assumed, but the likelihood
 37 consists of a combinatorial explosion of terms, and is therefore completely

1 intractable for all but trivial situations. Hence, in this example, the likeli-
2 hood function is simply not known. These examples may be considered toy
3 models in that one could easily maximize their associated objective functions
4 and compute sandwich matrices to obtain point estimates and asymptotic
5 confidence intervals. We use these examples simply to illustrate the effec-
6 tiveness of the OFS framework.

7 For each example in this section, we conduct a simulation study to in-
8 vestigate how well the OFS adjustment performs by measuring how often
9 nominal $(1 - \alpha)$ credible intervals cover θ_0 . To do this, we draw datasets
10 \mathbf{y}_k , $k = 1, \dots, 1000$, from the model determined by some fixed θ_0 . We
11 then run a random walk Metropolis algorithm, with $\ell_M(\theta; \mathbf{y}_k)$ inserted in
12 place of a likelihood, on each of the 1000 datasets. Next, we use each set
13 of MCMC samples to compute estimates $\hat{\theta}_{\text{QB},k}$ and $\hat{\Omega}_k$ using different es-
14 timators as discussed in Section 3.1. Finally, we use each $\hat{\theta}_{\text{QB},k}$ and $\hat{\Omega}_k$ to
15 adjust their corresponding batch of MCMC output, and record the resultant
16 equi-tailed $(1 - \alpha)$ credible interval for many values of α . In addition, we run
17 the curvature-adjusted sampler of Cooley et al. (2012) for comparison. For
18 each example, empirical coverage rates are plotted against nominal coverage
19 probabilities.

20 4.1 Tapered likelihood for spatial Gaussian processes

21 The most common structure for modeling spatial association among obser-
22 vations is the Gaussian process (Cressie, 1991; Stein, 1999). In addition to
23 modeling Gaussian responses, the Gaussian process has been used exten-
24 sively in hierarchical models to induce spatial correlation for a wide variety
25 of response types (Banerjee et al., 2004).

Here we assume that $Y(\mathbf{s}) \sim \text{GP}(0, C(\theta); \mathbf{s})$, a mean-zero Gaussian pro-
cess whose second-order stationary covariance is given by a parametric fam-
ily of functions C indexed by θ , depending on locations \mathbf{s} in some spatial
domain \mathcal{D} . We will further assume that the covariance between any two
observations y_i and y_j located at \mathbf{s}_i and \mathbf{s}_j is a function of only the distance
 $\|\mathbf{s}_i - \mathbf{s}_j\|$. Then the likelihood for n observations from a single realization of
 $Y(\mathbf{s})$ is

$$\begin{aligned} \ell_n(\theta; \mathbf{y}_n) = & -\frac{n}{2} \log(2\pi) - \frac{1}{2} \log(|\Sigma_n(\theta)|) \\ & - \frac{1}{2} \mathbf{y}_n' \Sigma_n(\theta)^{-1} \mathbf{y}_n, \end{aligned} \quad (8)$$

26 where $\Sigma_{i,j,n}(\theta) = C(\theta; \|\mathbf{s}_i - \mathbf{s}_j\|)$.

While conceptually simple, these Gaussian process models present computational difficulties when the number of observations of the Gaussian process becomes large, as the likelihood function (8) requires the inversion of a $n \times n$ matrix, which has computational cost $\mathcal{O}(n^3)$. To mitigate this cost, Kaufman et al. (2008) proposed replacing (8) with the *tapered* likelihood function

$$\begin{aligned} \ell_{t,n}(\boldsymbol{\theta}; \mathbf{y}_n) = & -\frac{n}{2} \log(2\pi) - \frac{1}{2} \log(|\boldsymbol{\Sigma}_n(\boldsymbol{\theta}) \circ \mathbf{T}_n|) \\ & - \frac{1}{2} \mathbf{y}_n' ((\boldsymbol{\Sigma}_n(\boldsymbol{\theta}) \circ \mathbf{T}_n)^{-1} \circ \mathbf{T}_n) \mathbf{y}_n, \end{aligned} \quad (9)$$

where the \circ notation denotes the element-wise product, and $\mathbf{T}_{ij} = \rho_t(\|\mathbf{s}_i - \mathbf{s}_j\|)$, a compactly-supported correlation function that takes a non-zero value when $\|\mathbf{s}_i - \mathbf{s}_j\|$ is less than some pre-specified “taper range.” The compact support of ρ_t induces sparsity in \mathbf{T}_n , and hence all operations required to compute (9) may be computed using specialized sparse-matrix algorithms, which are much faster and more memory-efficient than their dense-matrix analogues.

Under suitable conditions, the tapered likelihood satisfies asymptotics of the form (2), and Theorems 1 and 2 apply (Shaby and Ruppert, 2012). For the simulations, we take $C(\boldsymbol{\theta}; \|\mathbf{s}_i - \mathbf{s}_j\|) = \sigma^2 \exp\{-c/\sigma^2 \cdot \|\mathbf{s}_i - \mathbf{s}_j\|\}$, with $\boldsymbol{\theta} = (\sigma^2, c)' = (1, 0.2)'$. The observations are made on a 40×40 unit grid, so that each dataset \mathbf{y} is a single 1600-dimensional realization of a stochastic process. Half-Cauchy priors were used for both parameters.

For this example, analytical expressions for both $\mathbf{P}(\boldsymbol{\theta})$ and $\mathbf{Q}(\boldsymbol{\theta})$ are available (Shaby and Ruppert, 2012). As described in Section 3.1, we use the plug-in estimator $\hat{\Omega}_k = \hat{\mathbf{Q}}_{\text{III}}(\hat{\boldsymbol{\theta}}_{\text{QB},k})^{-1} \hat{\mathbf{P}}_{\text{II}}(\hat{\boldsymbol{\theta}}_{\text{QB},k})^{1/2} \hat{\mathbf{Q}}_{\text{III}}(\hat{\boldsymbol{\theta}}_{\text{QB},k})^{1/2}$, as well as $\hat{\Omega}_k = \hat{\mathbf{Q}}_{\text{I}}^{-1} \hat{\mathbf{P}}_{\text{II}}(\hat{\boldsymbol{\theta}}_{\text{QB},k})^{1/2} \hat{\mathbf{Q}}_{\text{I}}^{-1}$, with $\hat{\mathbf{Q}}_{\text{I}}^{-1}$ computed directly from the MCMC sample, for each $k = 1, \dots, 1000$ simulated datasets.

Figure 2 shows that the un-adjusted MCMC samples (dotted curves) yield horrible coverage properties for both σ^2 and c . It is somewhat interesting that while the “naive” intervals severely under-cover σ^2 , they severely over-cover c . We therefore see that a naive implementation results in being overly optimistic about estimates of σ^2 while being overly pessimistic about estimates of c . The OFS-adjusted intervals display much more accurate coverage, achieving nearly nominal rates, although for c , the asymptotic expression for $\hat{\mathbf{Q}}$ seems to produce intervals that are systematically slightly too short. The curvature-adjusted sampler results in similar coverage.

To explore how the marginal OFS adjustment differs from the conditional adjustment in the Gibbs sampler setting, we simulate data from a

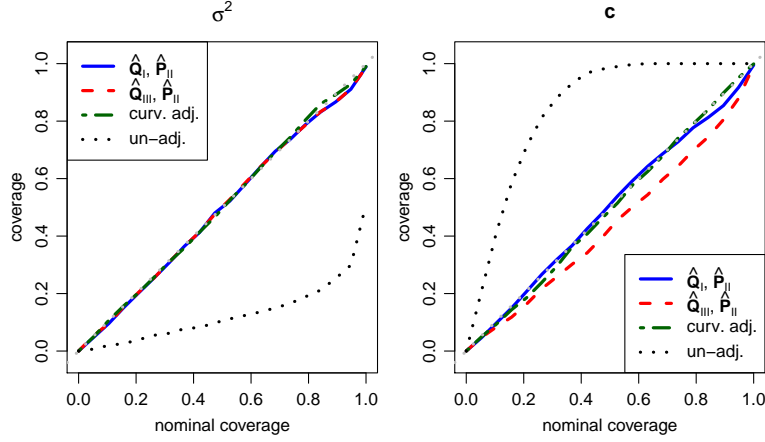


Figure 2: Empirical coverage rates for equi-tailed credible intervals based on MCMC samples using the tapered likelihood. Blue and red curves are OFS-adjusted samples using different estimates of Ω , green curves are from a curvature-adjusted sampler, and dotted curves are un-adjusted samples.

1 spatial linear model, $Y(\mathbf{s}) \sim \text{GP}(\mathbf{X}\boldsymbol{\beta}, C(\boldsymbol{\theta}); \mathbf{s})$. We set $\boldsymbol{\beta} = (-0.5, 0, 0.5)'$
2 and use the same spatial design and covariance function as above. The de-
3 sign matrix \mathbf{X} is a 1600×3 matrix of standard normal deviates, and the
4 prior distribution for $\boldsymbol{\beta}$ is a vague normal centered at zero. At each Gibbs
5 iteration, $\boldsymbol{\theta}$ is updated using a Metropolis step using the tapered likelihood,
6 and $\boldsymbol{\beta}$ is updated by drawing directly from its conditionally conjugate full
7 conditional distribution. The sampler is first run without adjustment, and
8 $\hat{\boldsymbol{\theta}}_{\text{QB}}$ and $\hat{\boldsymbol{\beta}}_{\text{QB}}$ are computed as the quasi-posterior means. The marginal
9 OFS adjustment matrix is then computed using the analytical expression
10 from Shaby and Ruppert (2012) by plugging in $\hat{\boldsymbol{\theta}}_{\text{QB}}$ and $\hat{\boldsymbol{\beta}}_{\text{QB}}$. Next, the
11 Gibbs sampler is run a second time using the estimated marginal OFS ma-
12 trix to adjust the sample from the full conditional distribution of $\boldsymbol{\theta}$ at each
13 iteration. Finally, the Gibbs sampler is run a third time, this time estimating
14 the conditional OFS adjustment matrix at each iteration by maximizing the
15 tapered likelihood function and plugging the resultant parameter estimates
16 into the asymptotic formula for $\hat{\Omega}$.

17 Because the conditional OFS-adjusted Gibbs sampler is so computa-
18 tionally expensive, we simulate just a few datasets and report the output
19 from one of them. Figures 3(a) and 3(b) compare the marginal adjusted
20 quasi-posterior distributions for the two covariance parameters, generated

1 with the marginal and conditional OFS adjustments. The qq-plot for σ^2
 2 shows almost perfect agreement except for a handful of MCMC samples on
 3 the upper tail. For the c parameter, the qq-plot shows that the marginal
 4 OFS adjustment produces a quasi-posterior that is the same shape as that
 5 of the conditional adjustment, but is slightly shifted to the right. Figure
 6 3(c) shows contours of a kernel density estimate of the joint marginal of-
 7 adjusted quasi-posterior distribution of the two covariance parameters, with
 8 the marginal adjustment in black and the conditional adjustment in gray.
 9 The contours are very similar, with some small differences appearing on
 10 the right half of the σ^2 -axis, indicating good agreement between the two
 11 bivariate distributions. The output from all the simulated datasets looked
 12 qualitatively similar, with no noticeable systematic differences between the
 13 two adjustments. The choice of adjustment had no discernible effect on the
 14 quasi-posterior distribution of β .

15 4.2 Composite likelihood for max-stable processes

16 Statistical models for extreme values that include spatial dependence are
 17 useful for studying, for example, extreme weather events like heat waves
 18 and powerful storms (Cooley et al., 2007; Sang and Gelfand, 2010, e.g.).
 19 Extreme value theory tells us that the distribution of block-wise maximum
 20 values (such as annual high temperatures) of independent draws from any
 21 distribution converges to a generalized extreme value (GEV) distribution
 22 (see Coles, 2001), if it converges at all. The asymptotics therefore suggest
 23 that any model of block-wise maxima at several spatial locations ought to
 24 have GEV marginal distributions with distribution function

$$F(y; \mu, \sigma, \xi) = \exp \left\{ - \left[1 + \xi \left(\frac{y - \mu}{\sigma} \right) \right]^{-1/\xi} \right\},$$

25 where μ is a location, σ a scale, and ξ a shape parameter that determines
 26 the thickness of the right tail. The GEV may be characterized by the max-
 27 stability property (Coles, 2001). More generally, block-wise maxima of ran-
 28 dom vectors must also converge to max-stable distributions.

29 Sang and Gelfand (2010) achieve spatial dependence with GEV marginals
 30 through a Gaussian copula construction. However, Gaussian copula models
 31 have been strongly criticized (Klüppelberg and Rootzén, 2006) because they
 32 do not result in max-stable finite-dimensional distributions, nor do they per-
 33 mit dependence in the most extreme values, referred to as *tail dependence*,
 34 and it is clear that physical phenomena of interest do exhibit strong spatial
 35 dependence even among the most extreme events.

1 An alternate approach for encoding spatial dependence of extreme values
2 is through max-stable process models (de Haan, 1984), which are stochastic
3 processes over some index set where all finite-dimensional distributions are
4 max-stable. Explicit specifications of spatial max-stable processes based on
5 the de Haan (1984) spectral representation have been proposed by Smith
6 (1990), Schlather (2002), and Kabluchko et al. (2009). These formulations
7 have the advantage that they do represent tail dependence.

8 Unfortunately, for all of the available spatial max-stable process models,
9 joint density functions of observations at three or more spatial locations are
10 not known (a slight exception is the Gaussian extreme value process (GEVP)
11 (Smith, 1990), for which Genton et al. (2011) derives trivariate densities).
12 Since bivariate densities are known, Padoan et al. (2010) proposes parameter
13 estimation and inference via the *pairwise likelihood*, where all bivariate log
14 likelihoods are summed as though they were independent:

$$\ell_p(\boldsymbol{\theta}; \mathbf{y}) = \sum_{i \neq j} \ell(y_i, y_j; \boldsymbol{\theta}).$$

15 The pairwise likelihood is a special case of a composite likelihood (Lindsay,
16 1988). Padoan et al. (2010) show that asymptotic normality of the form (2)
17 applies, so we may again apply the OFS adjustment.

18 Our simulation experiment consists of 1000 draws from a GEVP with
19 unit Fréchet margins on a 10×10 square grid, with 100 replicates per draw.
20 An example of a single replicate is shown in Figure 4. This setup would
21 correspond, for example, to 100 years of annual maximum temperature data
22 from 100 weather stations.

23 The unknown parameter $\boldsymbol{\theta}$ in a 2-dimensional GEVP is a 2×2 covariance
24 matrix. For this simulation, $\boldsymbol{\theta}_0 = (\Sigma_{11}, \Sigma_{12}, \Sigma_{22})' = (0.75, -0.5, 1.25)'$. The
25 prior distribution for $\boldsymbol{\Sigma}$ is a vague inverse-Wishart. For each draw, a long
26 MCMC chain is run and $\hat{\boldsymbol{\theta}}$ is computed as the posterior mean. In addition,
27 for each draw, all four $\mathbf{Q} - \mathbf{P}$ combinations of $\hat{\mathbf{Q}}_{\text{I}}$, $\hat{\mathbf{Q}}_{\text{II}}$, $\hat{\mathbf{P}}_{\text{I}}$, and $\hat{\mathbf{P}}_{\text{boot}}$, as
28 defined above, are computed to produce four estimates of $\boldsymbol{\Omega}$. Finally, the
29 curvature-adjusted MCMC sampler from Cooley et al. (2012) is run on each
30 simulated dataset, with $\boldsymbol{\Omega}$ estimated from $\hat{\mathbf{Q}}_{\text{II}}$ and $\hat{\mathbf{P}}_{\text{I}}$, evaluated at the
31 maximum pairwise likelihood estimate of $\boldsymbol{\Sigma}$.

32 Figure 5 shows that, for this simulation, the OFS adjustment produces
33 credible intervals that cover at almost exactly their nominal rates. Fur-
34 thermore, OFS-adjusted credible intervals based on the four values of $\hat{\boldsymbol{\Omega}}$
35 turned out nearly identical. This is about as good a result as one could
36 hope for. The curvature-adjusted sampler also achieves nominal coverage.

1 The un-adjusted intervals systematically under-cover for each of the three
2 parameters. This is expected (as noted by Cooley et al. (2012)), as the
3 pairwise likelihood over-uses the data by including each location in roughly
4 $n/2$ terms of the objective function rather than just one, as would be the
5 case with a likelihood function. This results in a pairwise likelihood surface
6 that is far too sharply-peaked relative to a likelihood surface. The OFS
7 adjustment seems to successfully compensate for this effect.

8 **5 Data analysis**

9 **5.1 Bird Counts**

10 Next, we apply tapered quasi-Bayesian analysis to a hierarchical model that
11 includes a Gaussian process. Instead of a purely spatial random field as
12 in Section 4.1, we assume a spatio-temporal random field, which highlights
13 some of the advantages of tapering over other methods designed for large
14 spatial datasets.

15 The dataset comes from a “citizen science” initiative called eBird (www.ebird.org). The idea of citizen science is that many non-professional ob-
16 servers can be leveraged to collect an enormous amount of data. eBird
17 participants across North America record the birds they see, along with
18 the time and location of the observation, into a web-based database. Here,
19 we look at 6114 observations of the Northern Cardinal in a section of the
20 eastern United States over a period from 2004 to 2007 (Figure 1).

22 Inspection of the data suggests an overdispersed Poisson model. Let
23 Y_i, \dots, Y_n be observed counts and $\mathbf{X} = [\mathbf{x}_1, \dots, \mathbf{x}_n]$ be a matrix of covariates
24 associated with each observation. Also, let $\mathbf{S} = [\mathbf{s}_1, \dots, \mathbf{s}_n]$ and $\mathbf{T} = [\mathbf{t}_1, \dots, \mathbf{t}_n]$
25 be the spatial and temporal locations, respectively, associated with Y_i, \dots, Y_n ,
26 with space indexed by latitude and longitude.

27 For this example, we deliberately chose a small number of predictors,
28 several of which vary spatially. Preliminary analyses led to a set of 10
29 covariates that includes time of day, day of year, human population den-
30 sity, percentage of developed open space (single-family houses, parks, golf
31 courses, etc.), tree canopy density, and variables that measure observer ef-
32 fort. Simple transformations (logs, powers, etc.) were applied to some of
33 the covariates, as suggested by ornithologists and exploratory analyses.

1 We specify the model as

$$\begin{aligned}
y_i | \lambda_i &\sim \text{Pois}(\lambda_i) \\
\log \lambda_i(\mathbf{x}_i, \mathbf{s}_i, \mathbf{t}_i) &= \mathbf{x}_i' \boldsymbol{\beta} + Z_i(\mathbf{s}_i, \mathbf{t}_i) + \varepsilon_i \\
\mathbf{z}(\mathbf{S}, \mathbf{T}) | \boldsymbol{\theta}^* &\sim N(0, \boldsymbol{\Sigma}^*(\boldsymbol{\theta}^*; \mathbf{S}, \mathbf{T})) \\
\varepsilon_i | \tau &\sim \text{iid } N(0, \sigma_\varepsilon^2).
\end{aligned} \tag{10}$$

2 We assume that the random effect $\mathbf{z}(\mathbf{S}, \mathbf{T})$ has a Gaussian random field
3 structure. Thus, this model is an example of “model-based geostatistics”
4 of Diggle et al. (1998), a class of spatial generalized linear models that has
5 seen wide application in the environmental literature.

6 Even though Northern Cardinals are not migratory birds, a spatio-
7 temporal structure for the random effect has a great intuitive appeal. One
8 can easily imagine clusters of birds habiting different locales, moving from
9 place to place based on things like food availability or disturbances. The
10 correlation range of the spatio-temporal random effect informs the ornitholo-
11 gists about the scales of movements of Northern Cardinals in space and time,
12 as well as providing clues about what un-measured covariates are needed to
13 explain the pattern of Northern Cardinal observations.

14 The parameter ε can be interpreted as either an overdispersion paramete-
15 ter, or as the traditional “nugget” effect, representing small-scale variation
16 or measurement error. It will be convenient to marginalize over the random
17 effects \mathbf{z} and ε and consider the distribution of the log-means directly. Fur-
18 thermore, we will write the matrix $\boldsymbol{\Sigma}^*(\boldsymbol{\theta}^*; \mathbf{S}, \mathbf{T}) + \sigma_\varepsilon^2 \mathbf{I}$ simply as $\boldsymbol{\Sigma}(\boldsymbol{\theta}; \mathbf{S}, \mathbf{T})$
19 and condense $\boldsymbol{\theta}^*$ and σ_ε^2 into the single parameter vector $\boldsymbol{\theta}$. The resulting
20 model, equivalent to (10), is written as

$$\begin{aligned}
y_i | \lambda_i &\sim \text{Pois}(\lambda_i) \\
\log \lambda_i(\mathbf{x}_i, \mathbf{s}_i, \mathbf{t}_i) &\equiv b_i \\
\mathbf{b}(\mathbf{X}, \mathbf{S}, \mathbf{T}) | \boldsymbol{\theta}, \boldsymbol{\beta} &\sim N(\mathbf{X}\boldsymbol{\beta}, \boldsymbol{\Sigma}(\boldsymbol{\theta}; \mathbf{S}, \mathbf{T})).
\end{aligned} \tag{11}$$

21 Another level in the hierarchy imposes a ridge penalty on the regression
22 coefficients $\boldsymbol{\beta}$, specified as

$$\boldsymbol{\beta} \sim N(0, \sigma_\beta^2 \mathbf{I}). \tag{12}$$

23 Finally, we need priors for the parameters $\boldsymbol{\theta}$ and σ_β^2

$$\begin{aligned}
\theta_i &\sim \pi_{\theta_i} \\
\sigma_\beta^2 &\sim \pi_{\sigma_\beta^2}
\end{aligned}$$

1 independently.

2 For $\Sigma(\boldsymbol{\theta}; \mathbf{S}, \mathbf{T})$, we chose a spatio-temporal covariance model from Gneit-
 3 ing (2002). The covariance functions described therein are nonseparable in
 4 that (except in special cases) they cannot be written as the product of a
 5 purely spatial and purely temporal covariance function. Specifically, we let

$$C(\boldsymbol{\theta}^*; \mathbf{h}, \mathbf{u}) = \frac{\sigma^2}{(a\mathbf{u}^{2\alpha} + 1)^2} \cdot \exp \left\{ \frac{-(c/\sigma^2)\mathbf{h}^{2\gamma}}{(a\mathbf{u}^{2\alpha} + 1)^{\omega\gamma}} \right\}, \quad (13)$$

6 where \mathbf{h} and \mathbf{u} are distances between observation points in space and time,
 7 respectively. The parameters $\alpha \in (0, 1]$ and $\gamma \in (0, 1]$ control the smoothness
 8 of the process. We fixed these parameters at convenient values of 1 and .5,
 9 respectively, because they were not well-identified by the data.

10 The parameter $\omega \in [0, 1]$ has the nice interpretation of specifying the
 11 degree of nonseparability between purely spatial and purely temporal com-
 12 ponents; when $\omega = 0$, $C(\boldsymbol{\theta}; \mathbf{h}, \mathbf{u})$ is the product of a purely temporal and a
 13 purely spatial (exponential) covariance function.

14 Priors for the parameters σ^2 , a , c , σ_ε^2 , and σ_β^2 are specified as vague
 15 Cauchy distributions, truncated to have only positive support. The interac-
 16 tion parameter ω is given a uniform prior on $[0, 1]$.

17 A valid spatio-temporal taper matrix may be constructed as the element-
 18 wise product of a spatial and a temporal taper matrix

$$\mathbf{T} = \mathbf{T}_s \circ \mathbf{T}_t.$$

19 Constructed this way, \mathbf{T} inherits the sparse entries of both \mathbf{T}_s and \mathbf{T}_t , and
 20 may therefore itself be extremely sparse.

21 Several other methods exist to mitigate the computational burden im-
 22 posed by large spatial datasets with non-Gaussian responses. Wikle (2002)
 23 and Royle and Wikle (2005) embed a continuous spatial process into a latent
 24 grid and work in the spectral domain using fast Fourier methods. However,
 25 applying Fourier methods here is problematic, as it is not obvious how to
 26 do so for a process that has spatio-temporal structures. Low rank meth-
 27 ods like predictive processes (Banerjee et al., 2008; Finley et al., 2009) and
 28 fixed rank Kriging (Cressie and Johannesson, 2008) are also popular for
 29 spatial data. Applying these methods to spatio-temporal models is possi-
 30 ble, but awkward. For predictive processes, one must decide how to specify
 31 knot locations in space \times time. For fixed rank Kriging, one must specify
 32 knot locations as well as space-time kernel functions. Fixed rank Kriging as
 33 been adapted to the spatio-temporal setting (Cressie et al., 2010) through a

1 linear filtering framework, but only for Gaussian responses. Finally, Gauss-
2 Markov approximations to continuous spatial processes are fast to compute,
3 especially when using Laplace approximations in place of MCMC (Lindgren
4 et al., 2011; Rue et al., 2009). However, again, these methods do not apply
5 to data with spatio-temporal random effects.

6 In contrast, application of the tapering approach in the spatio-temporal
7 context is immediate and even potentially enjoys increased computational
8 efficiency relative to the purely spatial context because of the additional
9 sparsity induced by element-wise multiplication with the temporal taper
10 matrix.

11 For the eBird data, a taper range of 20 miles and 60 days gives a tapered
12 covariance matrix with about .5% nonzero elements. MCMC was carried out
13 using a block Gibbs sampler. Each evaluation of the expensive normal log
14 likelihood was replaced by its tapered analogue. Within each Gibbs iteration,
15 each of \mathbf{b} , $\boldsymbol{\theta}$, and σ_{β}^2 are updated with a random walk Metropolis step.
16 The full conditional distribution for β is conditionally conjugate, enabling
17 a simple update as a draw from the appropriate normal distribution.

18 As described in Section 3.3, the tapered Gibbs sampler was run twice.
19 Samples from the first run were used to produce point estimates of $\boldsymbol{\theta}$ and
20 the marginal OFS adjustment matrix $\boldsymbol{\Omega}_{\boldsymbol{\theta}|\beta, \mathbf{b}, \sigma_{\beta}^2}$. The estimate $\hat{\boldsymbol{\Omega}}_{\boldsymbol{\theta}|\beta, \mathbf{b}, \sigma_{\beta}^2}$
21 was computed from the asymptotic expressions for \mathbf{Q} and \mathbf{Q} evaluated at
22 $\hat{\boldsymbol{\theta}}_{\text{QB}}$, the quasi-posterior mean. Because the quasi-posterior distribution of
23 interaction parameter ω was nearly uniform on $[0,1]$, it was excluded from
24 the adjustment. The conditional OFS was not attempted, as doing so would
25 have required several months of computation time.

26 After discarding 5000 burn-in iterations, 5000 MCMC samples were used
27 for estimation and prediction. Pointwise quantiles of the posterior correlation
28 surface are shown in Figure 6, for both the un-adjusted and adjusted
29 samples. The point at which the correlation drops to .05, often called the
30 “effective range” of a process, is the most extreme contour displayed in each
31 of the plots in figure 6. While the two sets of contours do not differ much
32 in the median, they are quite different in the upper and lower quantiles.
33 For this analysis, the correlation structure is a key component with a useful
34 interpretation, so its posterior uncertainty is of interest. Comparing the
35 OFS-adjusted and un-adjusted correlation surfaces, it is interesting to note
36 that OFS adjustment gives decreased temporal uncertainty but increased
37 spatial uncertainty.

38 The fairly long median effective range of around 225 days at spatial lag 0
39 seemed reasonable to a panel of ornithologists, as Northern Cardinals, while

1 they do move around to some degree, are not migratory birds. The effective
2 range of 3 miles at time lag 0 seemed reasonable as well. Northern Cardinals
3 build new nests each year and are socially monogamous within a breeding
4 season, but divorces sometimes occur between years. They generally stay
5 close to the nest to forage and bathe. Males are highly territorial and will
6 occasionally challenge neighboring males' breeding territories. These behav-
7 iors are consistent with a the posterior median temporal dependence range
8 of a significant fraction of a single year, and a posterior median spatial range
9 that is larger than but in the ballpark of an individual's territorial range.

10 Posterior estimates for some of the more interesting fixed effects, along
11 with 95% pointwise credible intervals, are plotted in Figure 7. The top
12 right panel shows a clearly increasing trend as a function of the number
13 of hours spent observing. The top right panel shows that the effect as a
14 function of time of day increases until about 8 a.m. and then decreases until
15 about 3 p.m., when it again begins to increase. The increase after 3 p.m.
16 is accompanied by very wide credible intervals. In the bottom left panel,
17 we can see an overwhelming negative effect at high elevations. Finally, the
18 bottom right panel shows a seasonal cycle that peaks in early winter and
19 attains its minimum in later summer.

20 Recall that these fixed effects are on the log scale. Here again, a panel
21 of ornithologists was pleased with the results. Obviously, the number of ob-
22 served counts should increase with the amount of time an observer spends
23 watching. The peak in the time of day effect at around 8 a.m. reflects the
24 time of the highest activity level of the birds. The wide confidence bands
25 starting at around 4 p.m. probably results from a lack of data in the after-
26 noon. Northern Cardinals cannot live in habitats found at higher elevations,
27 a fact reflected in the huge negative effect estimated after about 700 meters.
28 Finally, cardinals tend to be easier to detect during the winter months be-
29 cause they are more vocal, and they visit feeders more frequently. In the
30 summer months, they tend to stay more hidden because it is their breeding
31 season, and they do not visit feeders as often because food is more plentiful.
32 These seasonal variations in detectability are reflected in the pattern shown
33 in the estimated date effect.

34 The median posterior predicted surface (Figure 8) of the mean counts
35 was generated by drawing from the posterior predictive distribution at a
36 large set of sample points in the spatial domain, for fixed values of "effort"
37 covariates, and at a fixed time.

38 Maps like Figure 8, of course, vary in time as well as space. The most
39 prominent feature of the predicted surface is the very low values along the
40 Appalachians. This is a result of the huge elevation effect, and corroborates

1 expert knowledge. Another noticeable feature of the prediction surface is
2 the elevated counts around population centers. It is well-known among or-
3 nithologists that Northern Cardinals are most common in the suburbs. This
4 happens for two reasons. The first is that they are attracted to the many
5 bird feeders found in the suburbs. The second is that suburban habitat,
6 with landscaped gardens and mixes of open areas, shrubs, and trees, is ideal
7 habitat for cardinals.

8 **6 Discussion**

9 The open-faced sandwich adjustment provides a way to incorporate estimat-
10 ing functions that are not likelihoods into Bayesian-like models. While the
11 resulting inference does not enjoy the elegant formal probabilistic interpre-
12 tation of pure Bayesian analysis, it does inherit some of its most desirable
13 attributes: borrowing strength across parameters, the ability to work with
14 complicated hierarchical structures, and propagating uncertainty through-
15 out model components, to name a few. When the likelihood function is
16 unknown or has undesirable properties, the OFS adjustment allows one to
17 retain these beneficial features of Bayesian analysis while avoiding the need
18 to compute the likelihood function by substituting a suitable objective func-
19 tion in its place.

20 These benefits come with a few costs. First, the resultant MCMC sam-
21 ples may not be interpreted as though they came from a true Bayesian model.
22 Second, the coverage characteristics of the output are only as good as the
23 applicability of the asymptotic approximation and the practitioner’s ability
24 to estimate the sandwich matrix, which can be a difficult task in some sit-
25 uations (Kauermann and Carroll, 2001). Third, estimating the adjustment
26 matrix using sample moments or a bootstrap and relying on it to produce
27 posterior samples has a decidedly “un-Bayesian” feel to it. Finally, using the
28 adjustment in the Gibbs sampler context does require approximately twice
29 the computational effort as the un-adjusted Gibbs sampler, which in some
30 cases can be considerable.

31 In addition to these considerations, comparisons between the OFS ad-
32 justment and the curvature adjustment of Cooley et al. (2012) seem natural.
33 In our simulations, both adjustments performed extremely well. The cur-
34 vature adjustment shares with OFS both the advantages and disadvantages
35 described above. But in addition, the curvature adjustment, as implemented
36 in the data example in Cooley et al. (2012), has several additional drawbacks
37 to consider. First, Cooley et al. (2012) require an outside method to estimate

1 of θ and Ω , whereas the OFS adjustment uses the MCMC sample to esti-
2 mate θ and Ω . Using the un-adjusted quasi-posterior sample to estimate θ
3 and Ω as we do here takes advantage of borrowing strength, leveraging prior
4 information, etc. that simply maximizing $\ell_M(\theta)$ cannot. We note, however,
5 that this drawback in the curvature adjustment can easily be avoided. One
6 could easily apply the strategy that we suggest in Section 3.3, running the
7 sampler first without adjustment to estimate θ and Ω in a Bayesian-like
8 way, and then using these estimates to implement the curvature-adjusted
9 sampler. In the Metropolis context, this strategy requires twice the compu-
10 tational effort of OFS; since OFS is applied to the sample post hoc, there
11 is no need to run the sampler a second time. In the Gibbs sampler setting,
12 however, the computational burden is identical.

13 More obvious is the enormous computational cost imposed by estimating
14 the conditional adjustment matrix in the adaptive version of the curvature
15 adjusted Gibbs sampler favored by Cooley et al. (2012). This simply would
16 not have been feasible, for example, in the eBird example of Section 5.1. We
17 note that this complication can probably be avoided by using the marginal,
18 rather than the conditional, version of $\hat{\Omega}$. In fact, since their simulated
19 comparisons between the marginal and conditional forms of $\hat{\Omega}$ performed
20 so similarly, we are confused as to why Cooley et al. (2012) use the much
21 more computationally expensive conditional version in their data example.
22 In the end, conditional on implementation details, the OFS and curvature
23 adjustments are quite similar both in performance and in spirit.

24 Acknowledgements

25 This work was funded in part by NSF grants DMS-0914906, ITS-0612031,
26 ESI-0087760, and ONR grant N00244-11-1-009. The author wishes to thank
27 Cari Kaufman, Richard Smith, Dan Cooley, and Alan Gelfand for much
28 helpful discussion. Patient guidance on the eBird analysis was provided by
29 Wesley Hochachka, Steve Kelling, and especially Daniel Fink at the Cornell
30 Lab of Ornithology. Implementation of the GEVP was aided by generous
31 assistance from Mathieu Ribatet.

32 References

33 Sudipto Banerjee, Bradley P. Carlin, and Alan E. Gelfand. *Hierarchical*
34 *Modeling and Analysis for Spatial Data*. Monographs on Statistics and
35 Applied Probability. Chapman & Hall/CRC, Boca Raton, 2004.

- 1 Sudipto Banerjee, Alan E. Gelfand, Andrew O. Finley, and Huiyan Sang.
2 Gaussian predictive process models for large spatial data sets. *J. R. Stat.*
3 *Soc. Ser. B Stat. Methodol.*, 70(4):825–848, 2008.
- 4 M.J. Bayarri and J.O. Berger. The interplay of Bayesian and frequentist
5 analysis. *Statistical Science*, 19(1):58–80, 2004.
- 6 James O. Berger, Victor De Oliveira, and Bruno Sansó. Objective Bayesian
7 analysis of spatially correlated data. *J. Amer. Statist. Assoc.*, 96(456):
8 1361–1375, 2001.
- 9 V. P. Bhapkar. On a measure of efficiency of an estimating equation. *Sankhyā*
10 *Ser. A*, 34:467–472, 1972. ISSN 0581-572X.
- 11 Victor Chernozhukov and Han Hong. An MCMC approach to classical es-
12 timation. *J. Econometrics*, 115(2):293–346, 2003. ISSN 0304-4076.
- 13 Stuart Coles. *An introduction to statistical modeling of extreme values*.
14 Springer Series in Statistics. Springer-Verlag London Ltd., London, 2001.
15 ISBN 1-85233-459-2.
- 16 Daniel Cooley, Douglas Nychka, and Philippe Naveau. Bayesian spatial
17 modeling of extreme precipitation return levels. *J. Amer. Statist. Assoc.*,
18 102(479):824–840, 2007. ISSN 0162-1459.
- 19 Daniel Cooley, Mathieu Ribatet, and Anthony C. Davison. Bayesian infer-
20 ence from composite likelihoods, with an application to spatial extremes.
21 *Statistica Sinica*, 2012. doi: 10.5705/ss.2009.248.
- 22 N. Cressie and G. Johannesson. Fixed rank kriging for very large spatial
23 data sets. *J. Roy. Statist. Soc. Ser. B*, 70(1):209–226, 2008.
- 24 Noel Cressie, Tao Shi, and Emily L. Kang. Fixed rank filtering for spatio-
25 temporal data. *J. Comput. Graph. Statist.*, 19(3):724–745, 2010. ISSN
26 1061-8600. doi: 10.1198/jcgs.2010.09051. With supplementary material
27 available online.
- 28 Noel A. C. Cressie. *Statistics for spatial data*. Wiley Series in Probability and
29 Mathematical Statistics: Applied Probability and Statistics. John Wiley
30 & Sons Inc., New York, 1991. ISBN 0-471-84336-9. A Wiley-Interscience
31 Publication.
- 32 L. de Haan. A spectral representation for max-stable processes. *Ann.*
33 *Probab.*, 12(4):1194–1204, 1984. ISSN 0091-1798.

- 1 P. J. Diggle, J. A. Tawn, and R. A. Moyeed. Model-based geostatistics. *J.*
2 *Roy. Statist. Soc. Ser. C*, 47(3):299–350, 1998. ISSN 0035-9254. With
3 discussion and a reply by the authors.
- 4 David Draper. Coherence and calibration: comments on subjectivity and
5 “objectivity” in Bayesian analysis (comment on articles by Berger and by
6 Goldstein). *Bayesian Anal.*, 1(3):423–427 (electronic), 2006.
- 7 J. Durbin. Estimation of parameters in time-series regression models. *J.*
8 *Roy. Statist. Soc. Ser. B*, 22:139–153, 1960. ISSN 0035-9246.
- 9 Pedro E. Ferreira. Multiparametric estimating equations. *Ann. Inst. Statist.*
10 *Math.*, 34(3):423–431, 1982. ISSN 0020-3157.
- 11 A.O. Finley, H. Sang, S. Banerjee, and A.E. Gelfand. Improving the perfor-
12 mance of predictive process modeling for large datasets. *Computational*
13 *statistics & data analysis*, 53(8):2873–2884, 2009.
- 14 Reinhard Furrer, Marc G. Genton, and Douglas Nychka. Covariance taper-
15 ing for interpolation of large spatial datasets. *J. Comput. Graph. Statist.*,
16 15(3):502–523, 2006. ISSN 1061-8600.
- 17 Marc G. Genton, Yenyuan Ma, and Huiyan Sang. On the likelihood function
18 of Gaussian max-stable processes indexed by \mathbb{R}^d , $d \geq 1$. *Biometrika*, 2011.
19 to appear.
- 20 Tilmann Gneiting. Nonseparable, stationary covariance functions for space-
21 time data. *J. Amer. Statist. Assoc.*, 97(458):590–600, 2002. ISSN 0162-
22 1459.
- 23 V. P. Godambe and C. C. Heyde. Quasi-likelihood and optimal estimation.
24 *Internat. Statist. Rev.*, 55(3):231–244, 1987. ISSN 0306-7734.
- 25 Alastair R. Hall. *Generalized method of moments*. Advanced Texts in Econo-
26 metrics. Oxford University Press, Oxford, 2005. ISBN 0-19-877520-2.
- 27 Peter Hall, J. S. Marron, and Amnon Neeman. Geometric representation
28 of high dimension, low sample size data. *J. R. Stat. Soc. Ser. B Stat.*
29 *Methodol.*, 67(3):427–444, 2005. ISSN 1369-7412.
- 30 Lars Peter Hansen. Large sample properties of generalized method of mo-
31 ments estimators. *Econometrica*, 50(4):1029–1054, 1982. ISSN 0012-9682.
- 32 James W. Hardin and Joseph M. Hilbe. *Generalized estimating equations*.
33 Chapman & Hall/CRC, Boca Raton, FL, 2003. ISBN 1-58488-307-3.

- 1 Christopher C. Heyde. *Quasi-likelihood and its application*. Springer Series
2 in Statistics. Springer-Verlag, New York, 1997. ISBN 0-387-98225-6. A
3 general approach to optimal parameter estimation.
- 4 Peter J. Huber and Elvezio M. Ronchetti. *Robust statistics*. Wiley Series in
5 Probability and Statistics. John Wiley & Sons Inc., Hoboken, NJ, second
6 edition, 2009. ISBN 978-0-470-12990-6. doi: 10.1002/9780470434697.
- 7 Zakhar Kabluchko, Martin Schlather, and Laurens de Haan. Stationary
8 max-stable fields associated to negative definite functions. *Ann. Probab.*,
9 37(5):2042–2065, 2009. ISSN 0091-1798.
- 10 Göran Kauermann and Raymond J. Carroll. A note on the efficiency of
11 sandwich covariance matrix estimation. *J. Amer. Statist. Assoc.*, 96(456):
12 1387–1396, 2001. ISSN 0162-1459.
- 13 C. Kaufman, M. Schervish, and D. Nychka. Covariance tapering for
14 likelihood-based estimation in large spatial datasets. *J. Amer. Statist.*
15 *Assoc.*, 103(484):1545–1569, 2008.
- 16 John T. Kent. Robust properties of likelihood ratio tests. *Biometrika*, 69
17 (1):19–27, 1982. ISSN 0006-3444.
- 18 Claudia Klüppelberg and Holger Rootzén. Introduction to the copula dis-
19 cussion: some background. *Extremes*, 9(1):1–2, 2006. ISSN 1386-1999.
- 20 F. Lindgren, H. Rue, and J. Lindström. An explicit link between gaussian
21 fields and gaussian markov random fields: the stochastic partial differen-
22 tial equation approach. *J. Roy. Statist. Soc. Ser. B*, 73(4):423–498, 2011.
- 23 Bruce G. Lindsay. Composite likelihood methods. In *Statistical inference*
24 *from stochastic processes (Ithaca, NY, 1987)*, volume 80 of *Contemp.*
25 *Math.*, pages 221–239. Amer. Math. Soc., Providence, RI, 1988.
- 26 G.A.T. McVean, S.R. Myers, S. Hunt, P. Deloukas, D.R. Bentley, and
27 P. Donnelly. The fine-scale structure of recombination rate variation in
28 the human genome. *Science*, 304(5670):581, 2004.
- 29 R. Morton. Efficiency of estimating equations and the use of pivots.
30 *Biometrika*, 68(1):227–233, 1981. ISSN 0006-3444.
- 31 S.A. Padoan, M. Ribatet, and S.A. Sisson. Likelihood-based inference for
32 max-stable processes. *Journal of the American Statistical Association*,
33 105(489):263–277, 2010.

- 1 Christian P. Robert and George Casella. *Monte Carlo statistical methods*.
2 Springer Texts in Statistics. Springer-Verlag, New York, second edition,
3 2004. ISBN 0-387-21239-6.
- 4 J. Andrew Royle and Christopher K. Wikle. Efficient statistical mapping
5 of avian count data. *Environ. Ecol. Stat.*, 12(2):225–243, 2005. ISSN
6 1352-8505.
- 7 Håvard Rue, Sara Martino, and Nicolas Chopin. Approximate Bayesian
8 inference for latent Gaussian models by using integrated nested Laplace
9 approximations. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 71(2):319–392,
10 2009. ISSN 1369-7412. doi: 10.1111/j.1467-9868.2008.00700.x.
- 11 H. Sang and A.E. Gelfand. Continuous spatial process models for spatial
12 extreme values. *Journal of Agricultural, Biological, and Environmental*
13 *Statistics*, 15(1):49–65, 2010.
- 14 M. Schlather. Models for stationary max-stable random fields. *Extremes*, 5
15 (1):33–44, 2002.
- 16 Benjamin Shaby and David Ruppert. Tapered covariance: Bayesian estima-
17 tion and asymptotics. *J. Comp. Graph. Statist.*, 2012. in press.
- 18 Elizabeth L. Smith and Alec G. Stephenson. An extended Gaussian max-
19 stable process model for spatial extremes. *J. Statist. Plann. Inference*,
20 139(4):1266–1275, 2009. ISSN 0378-3758.
- 21 R.L. Smith. Max-stable processes and spatial extremes. *Unpublished*
22 *manuscript*, 1990.
- 23 Michael L. Stein. *Interpolation of spatial data*. Springer Series in Statistics.
24 Springer-Verlag, New York, 1999. ISBN 0-387-98629-4. Some theory for
25 Kriging.
- 26 L. Tian, J.S. Liu, and LJ Wei. Implementation of estimating function-based
27 inference procedures with Markov chain Monte Carlo samplers. *J. Amer.*
28 *Statist. Assoc.*, 102(479):881–888, 2007.
- 29 C. K. Wikle. Spatial modelling of count data: a case study in modelling
30 breeding bird survey data on large spatial domains. In *Spatial cluster*
31 *modelling*, pages 199–209. Chapman & Hall/CRC, Boca Raton, FL, 2002.
- 32 G. Yin. Bayesian generalized method of moments. *Bayesian Analysis*, 4(1):
33 1–17, 2009.

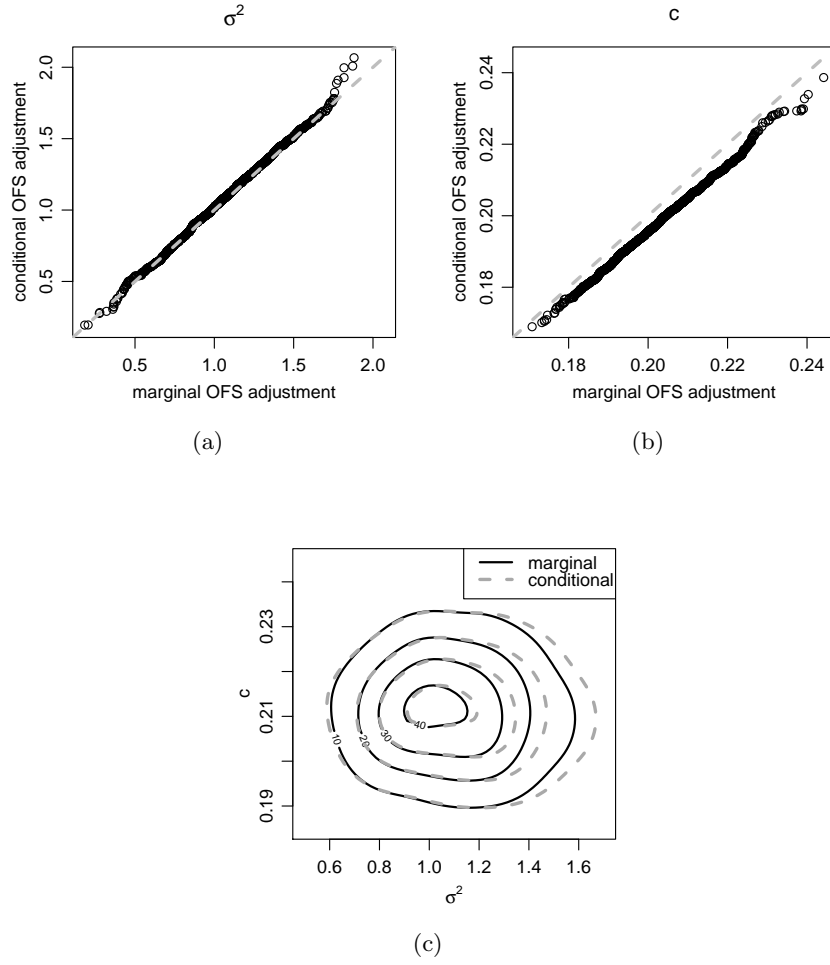


Figure 3: Comparison of marginal and conditional OFS-adjusted quasi-posterior distributions. Panels (a) and (b) show qq-plots of the marginal quasi-posteriors of the two covariance parameters. Panel (c) shows contour plots of a kernel density estimate of the joint quasi-posterior of the same parameters.

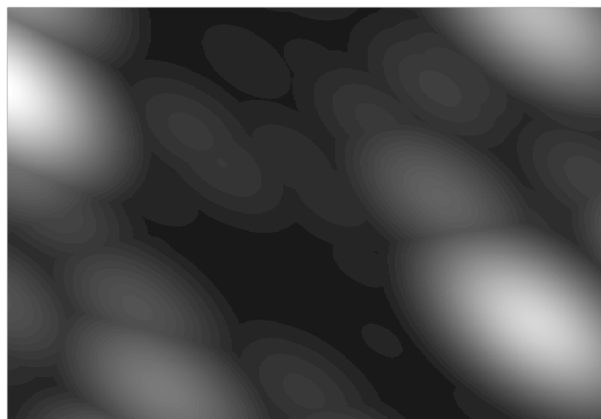


Figure 4: A realization of the Gaussian extreme value max-stable process of Smith (1990).

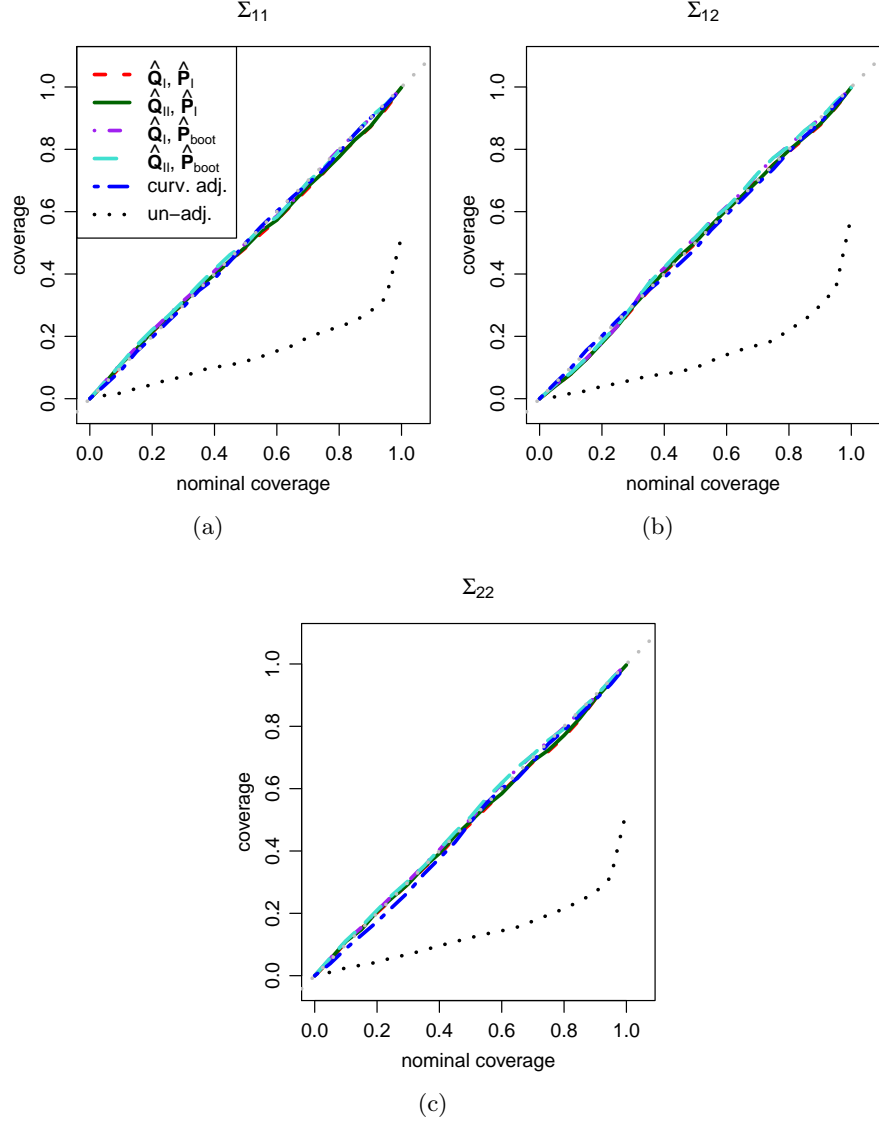


Figure 5: Empirical coverage rates for equi-tailed credible intervals based on MCMC samples using the pairwise likelihood. Colored curves are OFS-adjusted samples using different estimates of Ω , and from a curvature-adjusted sampler. Dotted curves are un-adjusted samples.

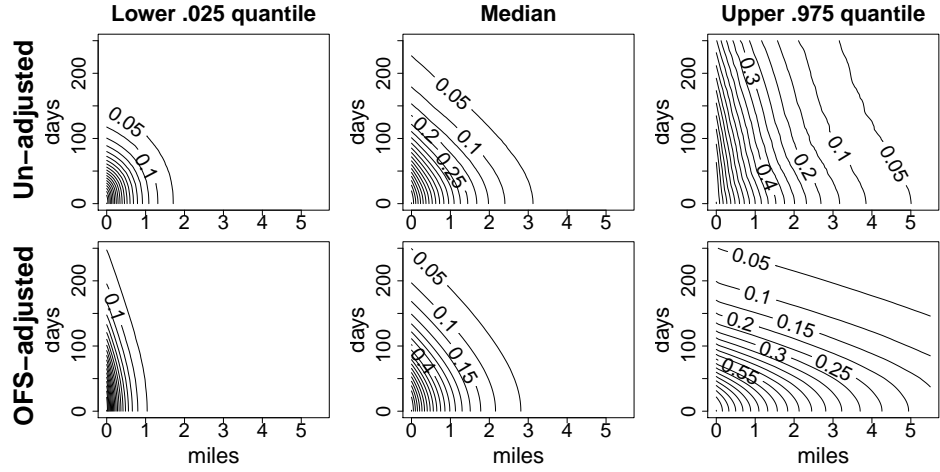


Figure 6: Posterior quantiles of spatio-temporal correlation surface. The top row was computed from the un-adjusted sampler, and the bottom row was computed from the OFS-adjusted sampler. The median surfaces are similar, but the high and low quantile contours show that the uncertainty is noticeably altered by the adjustment.

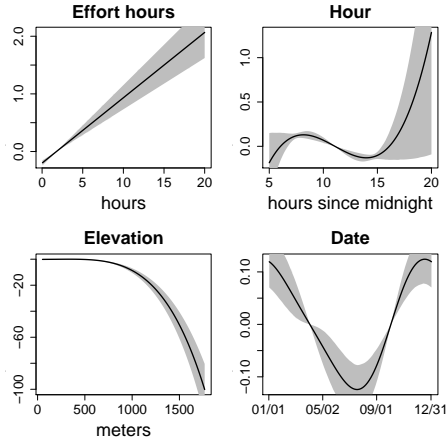


Figure 7: A subset of the estimated fixed effects. Solid lines are posterior medians, and shaded areas are posterior pointwise 90% credible intervals.

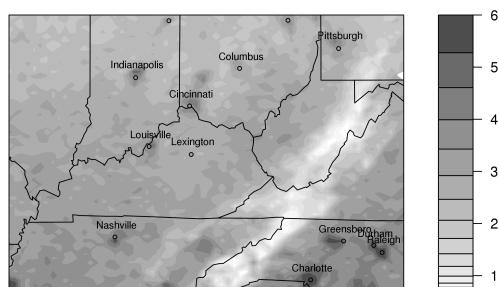


Figure 8: Median predicted surface for 8 a.m. on April 11.